Karlsruhe Institute of Technology

# Online Neural Network-based Language Identification

Master’s Thesis

Daniel H. Draper

Interactive Systems Lab - Institute for Anthropomatics and Robotics

# Contents

# Motivation

Karlsruhe Institute of Technology

- ASR and SLT commercially great success ["OK, Google", Siri, Jibbigo]
- Best performing ASR models are still trained on one source language.[1]

---

- Performance increase for multilingual tasks[1], UI streamlining.

# Motivation

Karlsruhe Institute of Technology

- ASR and SLT commercially great success [”OK, Google“, Siri, Jibbigo]
- Best performing ASR models are still trained on one source language.[1]

---

- Performance increase for multilingual tasks[1], UI streamlining.

# Motivation

- ASR and SLT commercially great success [”OK, Google“, Siri, Jibbigo]
- Best performing ASR models are still trained on one source language.[1]

---

- Performance increase for multilingual tasks[1], UI streamlining.

# Specific Applications

KIT
Karlsruhe Institute of Technology

## Lecture Translator

- Low latency, online application
- Small source language amount make performance increase likely

## European Parliament

- SLT already employed consistently[2]
- Many languages challenging, great potential.

# Specific Applications

## Lecture Translator

- Low latency, online application
- Small source language amount make performance increase likely

## European Parliament

- SLT already employed consistently[2]
- Many languages challenging, great potential.

# Specific Applications

## Lecture Translator

- Low latency, online application
- Small source language amount make performance increase likely

## European Parliament

- SLT already employed consistently[2]
- Many languages challenging, great potential.

# Specific Applications

Karlsruhe Institute of Technology

## Lecture Translator

- Low latency, online application
- Small source language amount make performance increase likely

## European Parliament

- SLT already employed consistently[2]
- Many languages challenging, great potential.

# Related Work I

## Historically

- PRLM[a], PPRM[b][3]
- Gaussian Mixture Models (GMM)s [4]

[a] single-language phone recognition followed by language-dependent, interpolated n-gram language modeling

[b] multiple single-language phone recognizers and language-dependent parallel phone recognition

# Related Work I

## Historically

- PRLM[a], PPRM[b][3]
- Gaussian Mixture Models (GMM)s [4]

[a] single-language phone recognition followed by language-dependent, interpolated n-gram language modeling

[b] multiple single-language phone recognizers and language-dependent parallel phone recognition

# Related Work II

## Modern Approaches

- VSM[a] + PPRM[5]
- I-Vector approaches[6][7]

[a] Vector Space Modelling

# Related Work II

## Modern Approaches

- VSM[a] + PPRM[5]
- I-Vector approaches[6][7]

[a] Vector Space Modelling

# Related Work III/Based on

## Similar Approaches

- Matejka et. al[8] similar setup: BNF $\rightarrow$ LID, with averaging 5 LID nets.
- [9] M. Heck et al. evaluate LID approaches prospect of Lecture Translator integration: PPRM/PRLM + Hybrid.

---

- Experimental baseline setup from Markus [10]: ”Language Adaptive DNNs for Improved Low Resource Speech Recognition“

# Related Work III/Based on

## Similar Approaches

- Matejka et. al[8] similar setup: BNF $\rightarrow$ LID, with averaging 5 LID nets.
- [9] M. Heck et al. evaluate LID approaches prospect of Lecture Translator integration: PPRM/PRLM + Hybrid.

---

- Experimental baseline setup from Markus [10]: "Language Adaptive DNNs for Improved Low Resource Speech Recognition"

# Related Work III/Based on

Karlsruhe Institute of Technology

## Similar Approaches

- Matejka et. al[8] similar setup: BNF $\rightarrow$ LID, with averaging 5 LID nets.
- [9] M. Heck et al. evaluate LID approaches prospect of Lecture Translator integration: PPRM/PRLM + Hybrid.

---

- Experimental baseline setup from Markus [10]: ”Language Adaptive DNNs for Improved Low Resource Speech Recognition“

# Theory of LID

- Given input $S = \{s(1), s(2), \ldots, s(T)\}$, $\{L_1, L_2, \ldots, L_N\}$

**Formulation of LID Task[11]**

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones $\upsilon$ (sequence $\Upsilon$):

$L^O = \arg\max_l P(\Upsilon|L_l)$

With Viterbi decoding on set of phone models $M$:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M)P(\Upsilon|L_l)}$$

# Theory of LID

- Given input $S = \{s(1), s(2), \ldots, s(T)\}$, $\{L_1, L_2, \ldots, L_N\}$

## Formulation of LID Task[11]

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones $\upsilon$ (sequence $\Upsilon$):

$L^O = \arg\max_l P(\Upsilon|L_l)$

With Viterbi decoding on set of phone models $M$:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M)P(\Upsilon|L_l)}$

# Theory of LID

- Given input $S = \{s(1), s(2), \ldots, s(T)\}$, $\{L_1, L_2, \ldots, L_N\}$

## Formulation of LID Task[11]

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones $\upsilon$ (sequence $\Upsilon$):

$L^O = \arg\max_l P(\Upsilon|L_l)$

With Viterbi decoding on set of phone models $M$:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M)P(\Upsilon|L_l)}$

# Theory of LID

- Given input $S = \{s(1), s(2), \ldots, s(T)\}$, $\{L_1, L_2, \ldots, L_N\}$

**Formulation of LID Task[11]**

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones $\upsilon$ (sequence $\Upsilon$):

$L^O = \arg\max_l P(\Upsilon|L_l)$

With Viterbi decoding on set of phone models $M$:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M)P(\Upsilon|L_l)}$$

# Theory of LID

- Given input $S = \{s(1), s(2), \ldots, s(T)\}$, $\{L_1, L_2, \ldots, L_N\}$

**Formulation of LID Task[11]**

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones $\upsilon$ (sequence $\Upsilon$):

$L^O = \arg\max_l P(\Upsilon|L_l)$

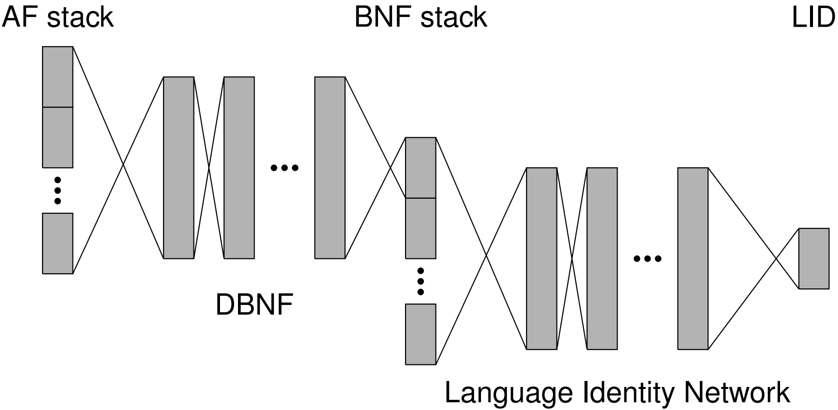
With Viterbi decoding on set of phone models $M$:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M)P(\Upsilon|L_l)}$$

# Experimental Setup



# Data Sets

## Euronews 2014

- 10 Languages (EN, FR, DE, AR, ES, IT, PO, PT, RU, TR)
- Original Corpus: 72h / Language, Reduced corpus: 18h / Language (Random Sampling of 10.000 speakers)
- 80 % train data, 10 % each dev/test set.

## Lecture Data

- 3 Languages (EN, FR, DE) 10h per language.
- KIT lectures, InterACT25, DGA talks
- Evaluation of EUNews results, Proof-of-concept for Lecture Translator integration.

# Data Sets

## Euronews 2014

- 10 Languages (EN, FR, DE, AR, ES, IT, PO, PT, RU, TR)
- Original Corpus: 72h / Language, Reduced corpus: 18h / Language (Random Sampling of 10.000 speakers)
- 80 % train data, 10 % each dev/test set.

## Lecture Data

- 3 Languages (EN, FR, DE) 10h per language.
- KIT lectures, InterACT25, DGA talks
- Evaluation of EUNews results, Proof-of-concept for Lecture Translator integration.

# Data Sets II

Karlsruhe Institute of Technology

## European Parliament

- 7 Languages (EN, FR, DE, ES, IT, PO, PT) 3.6h per language
- Simultaneous translation of all EUParliament speeches
- Further Evaluation, Proof-of-concept for EUParl integration.

# Feature Extraction

# DBNF

# DNN Structure

# Demo

# Evaluation Metrics

# Smoothing Filters

# Results

# Future Work

# Questions?

# References I


- Z. Tang, L. Li, and D. Wang, "Multi-task recurrent model for true multilingual speech recognition," *arXiv preprint arXiv:1609.08337*, 2016.
- "Technology and corpora for speech to speech translation," http://tcstar.org/.
- M. A. Zissman *et al.*, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing*, vol. 4, no. 1, p. 31, 1996.
- P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features." in *Interspeech*, 2002.

# References II


- H. Li, B. Ma, and C. H. Lee, “A vector space modeling approach to spoken language identification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271–284, Jan 2007.
- Y. Song, B. Jiang, Y. Bao, S. Wei, and L. R. Dai, “I-vector representation based on bottleneck features for language identification,” *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, November 2013.
- L. F. D’haro Enríquez, O. Glembek, O. Plchot, P. Matějka, M. Soufifar, R. d. Córdoba Herralde, and J. Černockỳ, “Phonotactic language recognition using i-vectors and phoneme posteriogram counts,” 2012.

# References III


P. Matejka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma, and B. Zhang, “Neural network bottleneck features for language identification,” *Proc. IEEE Odyssey*, pp. 299–304, 2014.

M. Heck, S. StÃ$\frac{1}{4}$ker, and A. Waibel, “A hybrid phonotactic language identification system with an svm back-end for simultaneous lecture translation,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4857–4860.

# References IV


M. MÃ¼ller, S. StÃ¼ker, and A. Waibel, “Language adaptive dnns for improved low resource speech recognition,” in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, USA, September 8-12 2016.

H. Li, B. Ma, and K. A. Lee, “Spoken language recognition: From fundamentals to practice,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.